Monitoring Report: 2009 Key Stage 2 Tests

Date: July 2010
Product code: Ofqual/10/4736

Contents

Monitoring of the 2009 Key Stage 2 test cycle

The monitoring undertaken by the Office of Qualifications and Examinations Regulation (Ofqual) in relation to the 2009 Key Stage 2 test cycle covered all aspects of the process. The exercises reported here centred on quality and timely delivery; we also monitored the logistics of printing the papers, their delivery to schools and the reporting of results. Over 1.7 million tests were taken by pupils. These then had to be marked within a six-week window and the results over three subjects delivered to schools on 7th July 2009. This was a major achievement, as the Chief Regulator, Kathleen Tattersall, said at the time:

> I am pleased that this year 99.9 per cent of results have been received by schools on time. Following the problems experienced last year, the timely delivery of results will be welcomed by schools, parents and pupils.
>
> As regulator, Ofqual is continuing to monitor the quality control of the marking of this year's papers, and we will be listening to schools about any concerns that they might have. Building on research already done by QC[D]A we will do some further work into the marking quality of this year's tests with the aim of reporting later this year.

We have now completed the additional work to which the Chief Regulator referred, and this report summarises our findings.

In early June 2009, the Qualifications and Curriculum Development Agency (QCDA) identified that there were 'quality assurance issues with English papers, but this is now being addressed and marking of these continues'.[1]

The 'quality assurance issues' recognised by QCDA in 2009 referred specifically to the marker training materials and process used for quality assurance of marking, which had led to more markers than expected being stopped. While QCDA chose to conduct quantitative research into these issues, which 'indicated that the most likely cause of the issues around marker standardisation were overly-ambitious marking tolerances' (see the Appendix, point 2 under 'Training markers and supervisors'), we chose to commission qualitative research of the English test papers and mark schemes, which was undertaken in the autumn of 2009 by consultants who had a mixture of subject expertise and Key Stage 2 experience. The research was specifically limited to the setting of the papers, their pre-testing in 2007

[1] Statement by QCDA spokesperson quoted in press reports, including the *Guardian*, 4th June 2009, www.guardian.co.uk/education/2009/jun/04/sats-markers-delays .

and 2008, the 2009 test cycle and the development and application of mark schemes.

We also monitored the marker training programme and level setting in 2009. This work was carried out by our staff. As scheduled, we focused on mathematics where the eight meetings of the training cascade were sampled. Additionally, monitors sampled the eight meetings for English and all but one for science. In total, 49 days of training and level setting were monitored.

The level setting process comprises three types of meeting: draft level setting, script scrutiny and final level setting. In 2009 our officers attended all three types of meeting across English, mathematics and science.

Monitors did not become involved in the meetings but responded to a predetermined list of questions designed to collect evidence of compliance with the code of practice and to record other observations.

Responsibilities and accountabilities

One of QCDA's key is responsibilities in delivering national curriculum tests is to ensure that an appropriate programme is in place for recruiting and training markers. In order to fulfil this responsibility, it appoints a test operations agency that must develop a training programme designed to train markers to apply the mark scheme consistently and in line with the agreed national standard. The training programme must also ensure that markers fully understand their roles and responsibilities and understand how to complete the necessary administration.

Context for the report

The process of developing tests starts two to three years before they are taken by pupils. The following summary is based on the QCDA document *Test Development: Level Setting and Maintaining Standards* (March 2010). The full document can be accessed on the QCDA website at www.qcda.gov.uk/resources/assets/QCDA_Assessments_test_setting.pdf .

Test development

Responsibility for the development and conduct of the assessments rests with QCDA. Each test is developed by an agency with appropriate expertise and is required to follow QCDA's specification.

'The specification ensures continuity from year to year and defines:

- the length of the test
- the coverage of the programme of study
- the characteristics of individual test questions
- the mark scheme requirements
- the balance of types of questions
- access to the test, including children with English as an additional language and children with special educational needs.' (p. 4)

Two pre-tests take place during the development. 'Prior to the pre-tests, the QCDA convenes test review group meetings and teacher panels to provide constructive feedback on the materials. Members of these groups are selected to provide a wide variety of educational experience and expertise so that feedback relates to the suitability and accessibility of the tests is as comprehensive as possible.' (p. 5)

'The first of the two pre-tests has the following purposes, it:

- determines how individual children respond to each question
- ensures that all children can understand the wording of each question and that they are not misunderstanding the question
- ensures that illustrations are appropriate and supportive
- obtains reliable data about the difficulty of individual questions.' (p. 6)

‘The first pre-test includes at least twice as many questions as are required for the final test.’ (p. 6)

Mark schemes are prepared and trialled alongside the tests to ensure that they properly reward pupils for their responses. The findings are discussed in meetings with teachers who have a wide range of educational experience and expertise, to check their suitability and accessibility. In the light of discussions, questions that are ambiguous or unfair to certain groups of children are modified or removed.

‘Since no substantial changes are possible after the second pre-test, it is important during this stage to ensure that data is fully understood and that the agency and QCDA are confident in their work putting together test booklets for the second pre-test. The test review group, which includes teachers and Key Stage experts, meets to look at the summaries of the statistical analyses and suggestions for amendments to questions in light of children’s responses.’ (p. 8)

So by the time of the second pre-test the test is in almost its final version. ‘The main purpose of the second pre-test is to obtain performance data about the tests as a whole in relation to the previous year’s test. … A sample of approximately 1,500 children takes the new test that is scheduled for use in schools the following year alongside the current year’s statutory test. In practice, the two tests are usually separated by about three or four weeks.’ (p. 9)

‘The principal means of equating the standard of the new test to the previous year’s test is to equate scores obtained by the sample of children taking the two tests.’ (p. 10) This information can be used alongside a detailed examination of scripts from both tests to identify scores on the new test that represent the same levels of performance as those on the statutory test. ‘The mark schemes are finalised against children’s responses to the second pre-test.’ (p. 11)

At this stage, the tests are handed over to QCDA for final checking before they are printed and distributed to schools.

Marker training

The markers are teachers, most with experience at Key Stage 2, who are trained nevertheless to ensure that they understand what is required and can apply the mark scheme accurately and consistently. ‘Marker training material is prepared by (the most senior marker) for each subject at each Key Stage, in close association with the test development agency, QCDA and test operations agency.’ (p. 12)

‘[The most senior marker] and test operations agency at this stage also use second pre-test data and experience of marking the second pre-test scripts to consider where they will advise QCDA to place tolerance bands determining the acceptability of markers’ work. These bands are known as absolute mark difference (AMD) bands’ and, since 2008, have been calculated by looking at the difference between the marks for the standardisation and benchmarking scripts awarded by a marker and the agreed marks awarded by the most senior marker for the same scripts.

‘Using the AMD, markers are placed into three bands (A, B and C). Markers in Band A will be the most accurate and consistent. Band C markers (there are very few each year) are not allowed to continue marking.’ (p. 12)

Level setting

‘Level setting is the process that determines the minimum number of marks needed to achieve a level. Threshold marks set for each subject must be in line with the national curriculum level description. This ensures standards are maintained and each pupil’s achievements are awarded the appropriate level.’ (p. 16)

‘The second pre-test involves children taking the new test alongside the statutory test. The data from both tests can be used to compare the relative demands between the two tests and establish the year-on-year continuity of standards. In January or February of each year, QCDA meets with test development agencies to set draft level thresholds’ (p. 13) ‘… to guide the script scrutiny process and inform the final level setting meeting.’ (p. 16)

‘Script scrutiny is the judgemental process by which performance on one test is compared to performance on another. This is to identify the scores on the second test that represent the same level of performance as that achieved on the first. The outcomes of script scrutiny are used alongside the pre-test data to inform the final level setting meeting.’ (p. 16)

Key successes of the 2009 cycle

Our staff attended training sessions in all three subjects tested at Key Stage 2 and found that meetings were well organised, with clear systems and procedures. Monitors noted three aspects that they wished to commend as good practice:

- a collegiate approach in which team members were encouraged to ask questions of the leaders to check their understanding of the mark scheme
- a professional approach by support staff, who understood their role in delivering and facilitating the discussions
- good training materials, including key messages about the tests and mark schemes, and helpful prompts, which could be used by supervisors in training meetings.

Overall, we found that QCDA and its agencies were compliant with the code of practice, although there were some minor differences in practice as noted later in this report.

Monitoring of the level setting meetings, through which standards are maintained from one year to the next, demonstrated that compliance with the requirements of the code of practice was very high. Overall the monitors were satisfied that standards were appropriate and were being maintained year-on-year.

Considering our findings in terms of five criteria common to all assessment schemes, we came to the following conclusions.

Validity

Our monitoring found the tests in mathematics, science and English to be valid for the statutory purpose of the tests. This purpose is defined ‘as ascertaining what they (pupils) have achieved in relation to the attainment targets for that (KS2) stage’ (Section 76 (1) Education Act 2002).

Reliability

The tests yielded results that were consistent across the country, irrespective of who marked the papers. The low number of changes to results after schools had requested reviews could be interpreted as evidence of the reliability of the tests. We monitored closely the quality assurance process. The meetings to train markers were generally effective in ensuring consistency.

Comparability

The level setting process worked well to ensure that standards in 2009 were equivalent to those in previous years.

Minimising bias

In mathematics and science, there were no issues reported of systematic bias. In English reading and writing there was no evidence of systematic bias, although some concerns are raised later in the report.

Manageability

For pupils, all the tests appear to have been appropriate in terms of length and complexity.

Developing effective tests and assessing the whole cohort of pupils at the end of the Key Stage is a massive exercise that requires the coordination of thousands of people. In 2009, this was achieved with considerable success and with the vast majority of results being delivered accurately and on time. This report, in concentrating on the issues identified, should not detract from that achievement.

Key concerns and areas where improvement is needed from the 2009 cycle

Complexity of the English mark scheme

During the early stages of the marker training cascade, we raised concerns with QCDA in relation to the lack of time available to deliver robust training for markers on both reading and writing. As a result, QCDA, after consultation with the senior marking teams, made some adjustments to the timings of training activities before, during and after the marker training day. Despite these mitigations, our monitors recorded that during marker training the time spent on training to mark writing was less than that spent on training to mark reading, and was rushed. In particular it was observed that the reading mark scheme appeared to be more complex than in previous years and so more time had to be spent addressing this issue.

Following the marking of the longer writing task at the pre-test stage, the test development agency recorded that 'it was noted that different valid approaches were used for report-writing about the trainers, including chronological styles. It will be important, therefore, to ensure that mark scheme wording and exemplar material accurately reflects the range (of) valid responses.'

However, the writing mark scheme did not explicitly make this point and on the marker training day the density of the reading training module resulted in less time than planned being given over to the writing training module. Consequently, this point may not have been impressed sufficiently upon the markers.

The qualitative research also expressed some concerns at the scale and complexity of the reading mark scheme. There were two main concerns. The first was the degree to which the pupils understood the demands of each question and what was likely to gain them a second or third mark where applicable. The second concern was the ability of the markers to deal with the range of alternative answers that were deemed more or less acceptable. Some questions were treated as if the answer was unequivocal when it was not. A lack of precision in framing more than half of the questions then forced the mark scheme to take account of alternative interpretations, leading to increased complexity and ambiguity.

The reading stimulus booklet, *No Place like Home,* comprised two sections. *Dear Norman* is a booklet containing a series of letters to a boy who has left home to live in his tree house in the back garden. *The Earthship* is a two-page

leaflet promoting an environmentally-friendly house built from recycled materials.

The following instances illustrate some of the concerns.

Question 15a asked about the humour in the situation. The element of differentiation is an overlay in the mark scheme. ‘Explain’ what else is funny about Norman’s situation encourages pupils to write about the comic aspects of his situation rather than to draw out incongruity or paradox as the mark scheme requires for two marks. That is one way of answering the question but another pupil, noting the availability of two marks, could have been inclined to describe two aspects. Pupils are also likely to find it funny that Norman stays in the tree house for only three days after leaving for good, but that is excluded from the indicative content in the mark scheme. The confusion in the mark scheme is apparent and it is difficult to distinguish between 2-mark and 1-mark answers. For example, ‘he’s left home but nobody seems that worried’ (2 marks) looks very similar to ‘the funny thing is his parents are letting him live there’ (1 mark); and ‘he hasn’t got out of school fully – he still gets homework sent’ (2 marks) looks very similar to ‘he’s not going to school and can do anything and eat anything he wants’ (1 mark).

A sample script provides an example of a reward of just 1 mark even though two humorous details are mentioned. It is a clear response to the question set: ‘He is living up a tree, writing and receiving letters from everyone he knows! He also enters competitions for newspapers while he’s up there!’

Examples of the complexity of the mark scheme include the following:

Question 15b necessitated three pages in the mark scheme for three marks, and 35 bulleted examples, while Question 24, over two pages, listed 21 acceptable points and two that were not acceptable.

Question 27 awarded three marks for explaining how Earthships could solve different types of problems. Greater detail in the mark scheme seemed simply to equate to longer length without providing clarity for the marker. The use of indicative content meant that there were six separate factors for markers to consider before arriving at a mark. While we recognise that the intention of the mark scheme was to increase accessibility to marks for pupils, it also increased the lack of manageability for markers.

Consistent practice and communication through the training cascade

Across the training programme the majority of supervisors and trainers appeared to be very content with the training. While it is not possible to

determine from observation alone whether or not the markers internalised fully the training or the effectiveness of the training programme, monitors noted a number of issues that may have impacted on the robustness and quality of training.

During the early phase of the marker training process, supervisors laid different degrees of emphasis on some aspects. At marker training meeting 4 in mathematics, some trainers adopted different approaches when delivering the training materials. Such inconsistencies may lead to confusion and to inconsistent application of the mark scheme as the marker training process cascades to less experienced markers. However, this particular concern was addressed by the provision of a training guide that all trainers were required to follow during subsequent training meetings.

Monitors observed that during marker training sessions, some marking personnel were not trained on all relevant questions. In English, in the later stages of the marker training cascade, apparently because of time pressure, some markers were not trained on straightforward reading questions.

Some references to other assessment models (single level tests) were also noted at training sessions in the early stages of the marker training cascade for mathematics, and these references were passed down the training cascade.

We are concerned that variation in training during the cascade, and references to other assessment models with different marking practices, may have led to dissemination of incomplete, confusing and inconsistent messages about the application of the mark scheme – particularly for new markers.

The code of practice precludes any changes being made to the mark scheme during training. However, changes to the additional guidance to supplement or correct the marking programme leaders' commentary and how such guidance should be applied to the mark scheme took place as late as the penultimate training event for Key Stage 2 mathematics. When coupled with the inconsistency in the way in which additional guidance was passed on by supervisors, this meant that marking personnel were unlikely to have received the same messages. In future, care should be taken to ensure that this does not jeopardise the quality and consistency of marking and make the published mark scheme – against which teachers will judge the quality of marking – inaccurate.

Errors in training materials and the quality of the scripts used in script scrutiny

Minor typographical errors were regularly found within the training materials throughout the training cascade. Supplementary/additional guidance was circulated to supervisory markers either orally or in writing – but not always in the same format to markers. We are concerned that errors in the training materials and corrections to them continued to be identified as late as the final marker training meeting for mathematics and English, and consequently put at risk the quality of the training of marking personnel.

In the script scrutiny exercises, while the level of compliance with the code of practice was high, a concern was raised that some of the scripts used in the meetings contained marking errors that would change the marks and therefore the pupil performance that they were supposed to represent. The concern relates not only to the lack of quality control allowing these scripts into script scrutiny, but also to the inconsistent manner with which they were dealt: one subject removed them, whereas another either corrected the mark, often without informing those who had already scrutinised the scripts, or just left the scripts at the now incorrect mark point.

Taking full account of minimising bias

The qualitative research regarded both reading texts as interesting and attractive, and felt that they offered opportunities for different kinds of response. However, there were some concerns over their use as part of a national test.

Dear Norman was linked to a particular culture, and the use of irony and sarcasm in this text, although part of the primary curriculum, might cause some children to be disadvantaged – particularly those for whom English is an additional language. However, evidence from the pre-testing process did not necessarily support this concern.

The Earthship was seen as an interesting mix of text, diagrams and illustrations, and the focus on recycling tied in well with the primary curriculum. The amount of information contained in the leaflet, and the use of technical language, might have made it challenging in the context of a timed test. At the second pre-test stage, pupils were asked the question ‘Did you enjoy *The Earthship* leaflet?’, to which 40 per cent responded negatively. Although there is no indication as to the reason for their response, it is fair to say that this is an unusually high number, and the pre-test report says that girls and pupils who were working at level 3 were more likely to respond negatively to the question.

The subject matter of the longer writing task in English was considered to be biased towards those pupils with an interest in sports activities, with possible access and equality issues for pupils with disabilities. Another aspect that caused concern was the amount of relatively technical vocabulary identified in the mark scheme as a performance discriminator, which may have disadvantaged those for whom English is an additional language. Examples included soles and insoles, laces, cross-over straps and moulding, and notions such as grip, bounce and friction. A diagram giving some of this vocabulary as part of the stimulus might have addressed this concern.

It was noted that after pre-testing, 30 per cent of teachers who responded were critical of the longer task, expressing concerns at the notion of writing a report and sustaining the writing at sufficient length. The fact that the mean score for this task was the lowest for any task set in the last four years lends added weight to the concerns raised. Although this factor would not have had an impact on overall standards, it might have had an impact on individual pupil performance.

Conclusions

We recognise that overall, the processes worked well and the results should be considered a fair reflection of each pupil's level of attainment. However, in a review such as this, there is an opportunity for improvements to be identified.

Validity

The English reading texts themselves were engaging and the open questions provided a good alternative to single response or multiple choice questions. However, the openness was potentially undermined by the complexity of the mark scheme.

The quality of responses to the questions obtained in the pre-tests could have been more carefully considered to identify what aspects genuinely differentiated different levels of response. The extensive analysis and discussion presented in pre-test reports could also have identified questions that would have been improved by rewording in order to signal the question setters' intentions more clearly.

For markers, the lack of guidance in the mark scheme – with regard to the range of different valid approaches that pupils might use for the longer task, and the lack of direction in the writing prompt to the pupils as to the form their writing should take – meant that it was difficult to judge responses in terms of their suitability to a particular genre. The shorter task was clear in setting out its purpose and form for pupils. It could have been improved, however, by more imaginative prompts and a clearer indication that succinct, descriptive writing was expected.

Reliability

Reliability relates to the propensity of an assessment procedure to generate consistent outcomes. This requires marking to be consistent across all those involved; this was largely true in mathematics.

In the reading test, the complexity of the mark scheme and the variations made to it through both oral and written additional guidance meant that it was difficult for consistency to be achieved. Although we accept that the most likely cause of the issue around marker standardisation were overly-ambitious marking tolerances, the difficulty of marking the reading test consistently may also have been reflected in this issue.

The use of indicative content to identify different levels of response led to mark schemes that were too extensive, and the distinctions to be made had to be inferred by the markers. The fact that up to 23 bullet points were deemed

necessary to cover allowed and disallowed answers in a 2-mark question suggests that the question itself could be either too broad in scope or flawed.

The problem of inconsistency was not confined to English. Errors were found in the training materials in other subjects, and differences in the ways in which those errors were addressed. The extensive use of additional guidance to supplement and even correct the mark schemes inevitably leads to inconsistency in marking and discrepancies between the marking and the published mark scheme released to teachers. QCDA should ensure that changes to the mark scheme are kept to a minimum, and that, when needed, clarifications are cascaded consistently to markers.

Comparability

Year-on-year comparability is an essential requirement of the national curriculum assessment exercise, and the level setting meetings are designed to ensure that the thresholds are based on statistical evidence from the pre-test and markers' judgement of pupil performance on the tests. Overall, the operation of the process was satisfactory and results across years can be compared.

The script scrutiny meetings were good but improvements could be made to the quality control procedures for the scripts used, to ensure that they contain no marking errors. Measures need to be taken to ensure that we can have confidence in both the quality of all the scripts presented at the script scrutiny meetings and the quality of discussion among all the scrutineers.

Minimising bias

It is essential that all materials and questions are checked for potential bias and full consideration given to all issues identified.

No problems were reported in science or mathematics. In English, the choice of reading resource material was rooted in a specific cultural context and relied on an ironic and mocking sense of humour. Together, these may have disadvantaged reluctant readers and pupils from other social, cultural and linguistic backgrounds. During the test development process, some members of the test review group felt that some of the letters required more than inference as children needed to tune into different characters and strategies that people use in life. While many test review group members gave positive feedback on the tests, some also felt that children with English as an additional language (EAL) might not understand the humour. Although the EAL analysis – which compares the performance of pupils who have English as an additional language with pupils who have English as a first language with an additional sample at the first pre-test – showed that the EAL sample was at no additional disadvantage, this may not be the case in every instance,

and concerns of this nature should be given careful consideration during the test development process in order to minimise bias.

The longer writing task was more obviously biased towards pupils with an interest in sports. Such pupils would be likely to have a greater specialist vocabulary, which would have benefited them in Composition and Effect. This bias was noted early in the development of the tests and should have been addressed.

Manageability

There is no evidence that pupils had problems dealing with the materials or producing their answers within the time allowed.

The main concerns centre on the manageability of the mark schemes for the markers and supervisors. In English, the reading mark scheme required very fine judgements to be made. The writing mark scheme, though broadly the same as in previous years, was not entirely secure with respect to viewpoint in the Composition and Effect strand.

The problems with standardisation led to further issues of manageability as some markers were stopped and others were required to submit further evidence. The setting and use of tolerances must be carefully monitored in future to ensure that the quality assurance process meets its purpose.

Recommendations and actions

Developing the tests and mark schemes

As a priority:

1. Reviews of test materials should involve the test development agency and the senior marking team of the test operations agency to ensure that the combination of question papers and mark scheme forms a sound basis on which the training and standardisation of markers can take place.

Additionally, QCDA should consider the following:

2. The commentaries on the standardisation and quality assurance scripts should be explicit and show how the marks have been awarded in order to support and extend the markers' understanding.

3. When expert reviewers within the test development cycle indicate that there may be issues of minimising bias against groups of pupils such as EAL or disabled pupils, these concerns should be seriously addressed and such questions given extra consideration of their place in a test.

Training markers and supervisors

As a priority:

1. QCDA should ensure that robust mechanisms for quality assuring materials are used in the marking process. Current methods of quality assuring the process of developing and delivering national curriculum assessments should be reviewed.

2. QCDA should investigate the impact that quality assurance exercises for English and science may have had on the quality of marking and the numbers of stopped markers.

Additionally, QCDA should consider the following:

3. QCDA needs to ensure that clear and consistent messages are disseminated to markers throughout the cascade.

4. The final version of the mark scheme should include all adjustments or clarifications. The efficacy of the mark scheme needs to be user tested by markers across a range of abilities before the commencement of the marker training process.

5. Particular consideration should be given to the structure and content of the training cascade for English, especially meetings 4, 6 and 8.

6. QCDA must ensure that there is clear communication of marking messages and requirements between the test operations agency and marking personnel. The interface between the test development agency and the test operations agency should be robustly managed by QCDA.

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2010

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk